PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-259089
(43)Date of publication of application : 03.10.1997

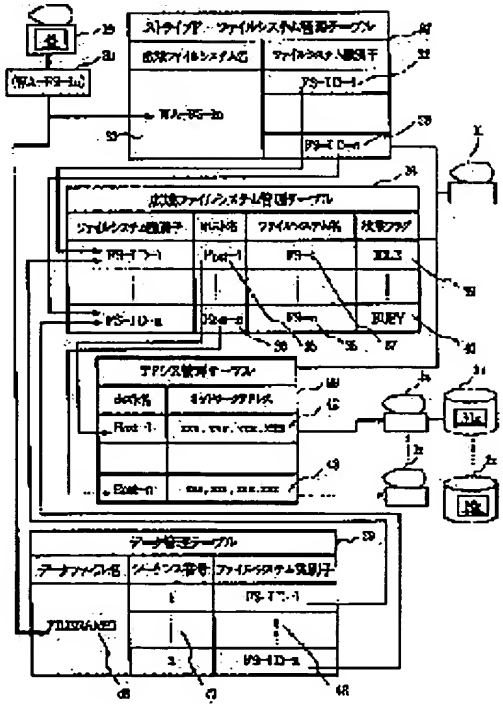
(51)Int.CI. G06F 15/16

(21)Application number : 08-071680
(22)Date of filing : 27.03.1996

(71)Applicant : NEC COMMUN SYST LTD
NEC CORP
(72)Inventor : OZAKI MITSUYOSHI
YOSHIDA KIYOHIKO

(54) DECENTRALIZEDLY NETWORKED STRIPED FILE SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To construct a striped file system with secondary storage devices connected to different host computers in decentralized network environment.
SOLUTION: Plural host computers are interconnected by a network and one of them is provided with a management table 26 for a wide-area file system, a management table 27 for the striped file system, a management table 28 for the network addresses of the host computers, and a data management table 29 for writing and reading data files divisionally. All the host computers while sharing and referring to those management tables write and read data to and out of secondary storage devices connected to a remote host computer, so that the data can be written to and read out of the secondary storage devices in parallel without increasing the loads on the host computers.



LEGAL STATUS

[Date of request for examination] 27.03.1996

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted

registration]

[Date of final disposal for application]

[Patent number]	2912221
[Date of registration]	09.04.1999
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	09.04.2003

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平9－259089

(43)公開日　平成９年(1997)10月３日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　15/16	３７０		Ｇ０６Ｆ　15/16	３７０Ｍ	

審査請求　有　　請求項の数３　ＯＬ　（全 18 頁）

(21)出願番号　特願平8－71680

(22)出願日　平成８年(1996)３月27日

(71)出願人　000232254
日本電気通信システム株式会社
東京都港区三田１丁目４番28号

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

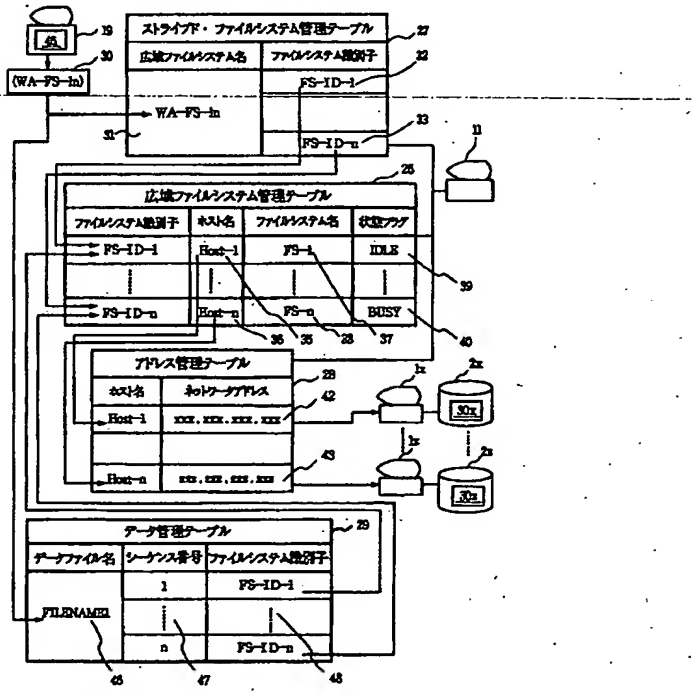
(72)発明者　小▲崎▼　光義
東京都港区芝五丁目７番１号　日本電気株式会社内

(72)発明者　▲吉▼田　清彦
東京都港区三田一丁目４番28号　日本電気通信システム株式会社内

(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】　分散ネットワーク化ストライプド・ファイルシステム

(57)【要約】
【課題】分散ネットワーク環境において、異なるホストコンピュータに接続された二次記憶装置からストライプド・ファイルシステムを構築可能とする。
【解決手段】複数のホストコンピュータをネットワークにより相互接続し、そのうちの１台に広域ファイルシステムの管理テーブル２６，ストライプド・ファイルシステムの管理テーブル２７，ホストコンピュータのネットワークアドレスの管理テーブル２８，及びデータファイルを分割して書き込み／読み出す為のデータ管理テーブル２９を配置する。全ホストコンピュータでこれらの管理テーブルを共有し参照しながら、ネットワークを介して遠隔のホストコンピュータに接続されている二次記憶装置に対するデータの書き込み／読み出しを行うことにより、ホストコンピュータの負荷を増加させることなく、複数の二次記憶装置に対するデータの並列書き込み／並列読み出しを可能とする。



1

【特許請求の範囲】

【請求項１】　ホストコンピュータに接続された複数の二次記憶装置上にそれぞれ構築された複数のファイルシステムを仮想的な１つのファイルシステムとし、この仮想的なファイルシステムに対するデータの書き込み及び読み出し処理をソフトウェアにより前記二次記憶装置単位に分割し並列分散化して行うストライプド・ファイルシステムにおいて、

前記複数の二次記憶装置の各々を互いに重複することなく接続した複数の前記ホストコンピュータを設け、これらホストコンピュータ間を相互接続して前記ホストコンピュータ及び対応する前記二次記憶装置上に構築された前記ファイルシステムを構成単位とするコンピュータネットワークを構築し、前記ホストコンピュータに接続された前記二次記憶装置上に構築された前記ファイルシステムの各々に対して仮想的なファイルシステム名を付与し、前記コンピュータネットワークを構成する前記ホストコンピュータから前記二次記憶装置に対するデータの入出力を前記仮想的なファイルシステム名を経由して行うことを特徴とする分散ネットワーク化ストライプド・ファイルシステム。

【請求項２】　前記ホストコンピュータの障害により前記二次記憶装置に対する分割されたデータの書き込みが失敗すると、代替のホストコンピュータを選別し、この代替のホストコンピュータに対応する前記二次記憶装置に対して該当データの書き込みを行うことを特徴とする請求項１記載の分散ネットワーク化ストライプド・ファイルシステム。

【請求項３】　広域ファイルシステム管理用ホストコンピュータに接続されている前記二次記憶装置上に配置された広域ファイルシステムを管理する為の広域ファイルシステム管理テーブルと、前記広域ファイルシステムから構成されるストライプド・ファイルシステムを管理する為のストライプド・ファイルシステム管理テーブルと、相互接続された前記ホストコンピュータを一意に識別する為のネットワークアドレス管理テーブルと、複数の前記二次記憶装置に分割されて格納されているデータファイルを管理する為のデータ管理テーブルとを備え、前記広域ファイルシステム管理テーブル，ストライプド・ファイルシステム管理テーブル，ネットワークアドレス管理テーブル及びデータ管理テーブルの情報を相互接続された全ての前記ホストコンピュータで共有して参照しながら前記コンピュータネットワークを通して遠隔の前記ホストコンピュータと通信を行うことによって、１つのデータファイルを分割して複数の前記ホストコンピュータに接続されている前記二次記憶装置に転送することを特徴とする請求項１または２記載の分散ネットワーク化ストライプド・ファイルシステム。

【発明の詳細な説明】

【０００１】

2

【発明の属する技術分野】本発明は分散ネットワーク化ストライプド・ファイルシステムに関し、特にホストコンピュータに接続された複数の二次記憶装置上にそれぞれ構築された複数のファイルシステムを仮想的な１つのファイルシステムとし、この仮想的なファイルシステムに対するデータの書き込み及び読み出し処理をソフトウェアにより前記二次記憶装置単位に分割し並列分散化して行うストライプド・ファイルシステムに関する。

【０００２】

【従来の技術】従来から、プロセス当たりのディスクに対するデータの入出力を行う際の性能を向上させる為の手段として、図１７に示すように、１台のホストコンピュータ８０に接続された複数の磁気ディスク装置のような二次記憶装置８１，８２，８３，８４上に構築したファイルシステムから、仮想的な１つのファイルシステム８５を構築し、この仮想的なファイルシステムを構成する各々の二次記憶装置の入出力制御装置に対してプロセスが並列にデータの入出力要求を行う、いわゆるストライプド・ファイルシステムの手法が採用されている。従来のストライプド・ファイルシステムの手法では、図１８に示すように、１台のホストコンピュータ８６の異なる入出力制御装置８７，８８に接続された複数の磁気ディスク装置のような二次記憶装置８９，９０上に構築されたファイルシステムを基本単位としてストライプド・ファイルシステム９１を構築し、各二次記憶装置上に構築されたファイルシステムに対する入出力処理を複数の入出力制御装置８７，８８に分散させ並列に行うことにより、プロセス当たりの二次記憶装置に対する入出力性能の向上を実現していた。

【０００３】

【発明が解決しようとする課題】この従来のストライプド・ファイルシステムの手法では、単一プロセス当たりの二次記憶装置に対するデータの入出力時の性能向上を図ることは出来たが、システム全体で見た場合、統合的なデータの入出力時の性能向上には至らなかった。すなわち、従来技術では、１台のホストコンピュータ８０に接続された複数台の二次記憶装置８１～８４上に構築されたファイルシステムを構築単位としていた。従って、ストライプド・ファイルシステム８５を構成する複数台の二次記憶装置８１～８４上のファイルシステムに対するデータの入出力処理を全て１台のホストコンピュータ８０で処理しなければならなかったため、ストライプド・ファイルシステムを構成する個々の二次記憶装置に対するデータの入出力処理を並列に行っても、一つのプロセスがデータの入出力に要する時間は並列化により短縮することは可能であるが、ホストコンピュータが処理しなければならない入出力データの量は変わらないため、二次記憶装置に対するデータの入出力時にホストコンピュータにかかる負荷を減少させることはできず、ホストコンピュータ全体の性能を向上させることはできなかっ

た。

【０００４】また、従来のストライプド・ファイルシステムの手法では、プロセス当たりの二次記憶装置に対するデータの入出力時の性能向上を図る為には、異なる入出力制御装置に接続された二次記憶装置上に構築されたファイルシステムを基本単位としてストライプド・ファイルシステムを構築しなければならないというハードウェア上の制約があった。すなわち、図１９に示すように、複数台の二次記憶装置に対するデータの入出力を１台の入出力制御装置９３で行った場合は、ストライプド・ファイルシステム９６に対する入出力処理を並列化しても、実際の二次記憶装置９４，９５へのデータの入出力処理は１台の入出力制御装置９３によってシーケンシャルに行われる為、単一プロセス当たりの二次記憶装置に対する入出力時の性能向上を図ることができなかった。従って、ホストコンピュータに搭載されている入出力制御装置の数によっては、ストライプド・ファイルシステムを構築してもプロセス当たりの二次記憶装置に対する入出力時の性能が向上されないという問題点があった。

【０００５】さらに、従来のストライプド・ファイルシステムの手法では、構築できるストライプド・ファイルシステムの大きさは、１台のホストコンピュータに接続可能な二次記憶装置の数により制限されていた。その理由は、１台のホストコンピュータに２台以上の磁気ディスク装置のような二次記憶装置を接続し、各々の二次記憶装置上に構築したファイルシステムを、ソフトウェアからの制御により仮想的に１つのファイルシステムに見せかけて、その仮想的なファイルシステムに対してデータの入出力を行っていたためである。

【０００６】したがって本発明の目的は、ホストコンピュータを二次記憶装置ごとに設け、これらホストコンピュータ間を相互接続してコンピュータネットワーク（分散ネットワーク）を構築し、各二次記憶装置にまたがった仮想的なファイルシステム（ストライプド・ファイルシステム）に対してネットワークを介していることを意識しないで分散・並列化させてデータの入出力を行うことにより、スループットの低下を招くこと無くストライプド・ファイルシステムに対するデータの入出力を実現すると共に、プロセス当たりの二次記憶装置に対する性能の向上だけでなく、システム全体を通じての二次記憶装置に対する性能を向上させ、従来よりも大きな容量を持つファイルシステムを実現することができるような分散ネットワーク化ストライプド・ファイルシステムを提供することにある。

【０００７】

【課題を解決するための手段】本発明の分散ネットワーク化ストライプド・ファイルシステムは、ホストコンピュータに接続された複数の二次記憶装置上にそれぞれ構築された複数のファイルシステムを仮想的な１つのファイルシステムとし、この仮想的なファイルシステムに対するデータの書き込み及び読み出し処理をソフトウェアにより前記二次記憶装置単位に分割し並列分散化して行うストライプド・ファイルシステムにおいて、前記複数の二次記憶装置の各々を互いに重複することなく接続した複数の前記ホストコンピュータを設け、これらホストコンピュータ間を相互接続して前記ホストコンピュータ及び対応する前記二次記憶装置上に構築された前記ファイルシステムを構成単位とするコンピュータネットワークを構築し、前記ホストコンピュータに接続された前記二次記憶装置上に構築された前記ファイルシステムの各々に対して仮想的なファイルシステム名を付与し、前記コンピュータネットワークを構成する前記ホストコンピュータから前記二次記憶装置に対するデータの入出力を前記仮想的なファイルシステム名を経由して行う構成を有する。

【０００８】また、上記構成において、前記ホストコンピュータの障害により前記二次記憶装置に対する分割されたデータの書き込みが失敗すると、代替のホストコンピュータを選別し、この代替のホストコンピュータに対応する前記二次記憶装置に対して該当データの書き込みを行う構成とすることができる。

【０００９】さらに、広域ファイルシステム管理用ホストコンピュータに接続されている前記二次記憶装置上に配置された広域ファイルシステムを管理する為の広域ファイルシステム管理テーブルと、前記広域ファイルシステムから構成されるストライプド・ファイルシステムを管理する為のストライプド・ファイルシステム管理テーブルと、相互接続された前記ホストコンピュータを一意に識別する為のネットワークアドレス管理テーブルと、複数の前記二次記憶装置に分割されて格納されているデータファイルを管理する為のデータ管理テーブルとを備え、前記広域ファイルシステム管理テーブル，ストライプド・ファイルシステム管理テーブル，ネットワークアドレス管理テーブル及びデータ管理テーブルの情報を相互接続された全ての前記ホストコンピュータで共有して参照しながら前記コンピュータネットワークを通して遠隔の前記ホストコンピュータと通信を行うことによって、１つのデータファイルを分割して複数の前記ホストコンピュータに接続されている前記二次記憶装置に転送する構成とすることができる。

【００１０】次に本発明の作用を説明する。複数台のホストコンピュータを相互に接続して構築したコンピュータネットワークの環境（分散ネットワーキング環境）において、ホストコンピュータに接続されている磁気ディスク装置のような二次記憶装置に、ネットワークを介していることを意識しないでデータの入出力を行うことができるようなファイルシステム（広域ファイルシステム）を構築し、異なるホストコンピュータに接続された二次記憶装置上に構築された広域ファイルシステムを構

成単位とするストライプド・ファイルシステムを構築する機能を提供することにより、ストライプド・ファイルシステムに対するデータの入出力時にホストコンピュータにかかる負荷を複数のホストコンピュータに搭載された入出力制御装置に分散させることで、スループットの低下を招くこと無くストライプド・ファイルシステムに対するデータの入出力を実現すると共に、ストライプド・ファイルシステムに対するデータの入出力処理そのものを、ネットワーク内の他のホストコンピュータに分散・並列化させて実行させることにより、プロセス当たりの二次記憶装置に対する性能の向上だけでなく、システム全体を通じての二次記憶装置に対する性能を向上させ、更に、ネットワーク内の複数のホストコンピュータに接続された二次記憶装置にまたがった仮想的なファイルシステムを構築することにより、従来よりも大きな容量を持つファイルシステムを実現することができる。

【００１１】また、プロセスがストライプド・ファイルシステムに対してデータの入出力を行うと、各ホストコンピュータ上で動作しているソフトウェアが、広域ファイルシステムを管理するためのホストコンピュータに接続されている二次記憶装置上の広域ファイルシステムを管理する為の管理テーブル，広域ファイルシステムから構成されるストライプド・ファイルシステムを管理する為の管理テーブル，相互接続されたホストコンピュータを一意に識別する為のネットワークアドレス管理テーブル，及びデータファイルを管理する為のデータ管理テーブルを参照し、プロセスによってデータの入出力が行われたストライプド・ファイルシステムを構成しているコンピュータネットワーク内のホストコンピュータに接続された二次記憶装置上に構築されたファイルシステムを識別し、データの分割及び分割されたデータの再構築、ならびにデータの入出力制御の分散・並列化を自動的に行う。このため、プロセスはストライプド・ファイルシステムに対してデータの入出力を行うだけで、データの断片化及び入出力処理の並列化、ならびにコンピュータネットワークで相互に接続された遠隔のホストコンピュータとの間でのデータの送受信を意識することなく、遠隔のホストコンピュータに接続された二次記憶装置に対するデータの入出力を実現することができる。

【００１２】

【発明の実施の形態】本発明の実施の形態について図面を参照して説明する。

【００１３】図１は本発明の第１の実施形態例を示す概略システム構成図である。図１において、本発明を適用するファイルシステムは広域ファイルシステムを成し、広域ファイルシステムを管理するためのホストコンピュータ１１と、その管理下の複数のホストコンピュータ１２，１３，１４，…，１ｍ、これら各ホストコンピュータ１１～１ｍにそれぞれ接続されている二次記憶装置２１，２２，２３，２４，…，２ｍと、各ホストコンピュータ１１～１ｍ間を相互接続しているネットワーク１０とから構成される。各ホストコンピュータ１２～１ｍに接続されている二次記憶装置２２～２ｍ上に構築されたファイルシステムから仮想的なストライプド・ファイルシステム２０が構築されている。ネットワーク１０には広域ファイルシステムを利用するホストコンピュータ（図示せず）が接続される。広域ファイルシステムを管理するためのホストコンピュータ１１に接続されている二次記憶装置２１には、図２に示すように、広域ファイルシステムを管理するための管理テーブル２６と、ストライプド・ファイルシステムを管理するための管理テーブル２７と、ホストコンピュータのホスト名とネットワーク内でホストコンピュータを一意に識別するためのアドレス情報を管理するためのアドレス管理テーブル２８と、データを複数の二次記憶装置に分割して書き込む際、及び複数の二次記憶装置に分割されて書き込まれているデータを復元する際に使用するデータ管理テーブル２９とが設けられている。

【００１４】広域ファイルシステムを管理するための管理テーブル２６は、図３に示すように、ネットワークを構成するホストコンピュータに対して付けられたホスト名（例えば、Ｈｏｓｔ－１，…）を登録する領域２６１と、そのホスト名のホストコンピュータに接続された二次記憶装置上に構築されたファイルシステムのファイルシステム名（例えば、ＦＳ－１，…）を登録する領域２６２と、ホスト名及びファイルシステム名の組に対して一意に与えられるファイルシステム識別子（例えば、ＦＳ－ＩＤ－１，…）を登録する領域２６３と、そのファイルシステム名のファイルシステムの状態を表す状態フラグを登録する領域２６４とから構成される。状態フラグは、ＢＵＳＹ状態（ファイルシステムに対する入出力が行われている状態）及びＩＤＬＥ状態（ファイルシステムに対する入出力が行われていない状態）の２つの状態を有する。

【００１５】ストライプド・ファイルシステム管理テーブル２７は、図４に示すように、各ホストコンピュータが広域ファイルシステム上のファイルにアクセスする際に参照する広域ファイルシステム名（例えば、ＷＡ－ＦＳ－１ｎ，…）を登録する領域２７１と、その広域ファイルシステム名に対応するファイルシステムを構成する複数のファイルシステムのファイルシステム識別子（例えば、ＦＳ－ＩＤ－１１，…）を登録する領域２７２とから構成される。

【００１６】アドレス管理テーブル２８は、図５に示すように、ホストコンピュータに対して付けられたホスト名（例えば、Ｈｏｓｔ－１，…）を登録する領域２８１と、ネットワーク内でホストコンピュータを一意に識別するためのネットワークアドレス情報（例えば、ａａａ．ａａａ．ａａａ．ａａａ，…）を登録する領域２８２とから構成される。

【００１７】データ管理テーブル２９は、図６に示すように、ストライプド・ファイルシステムにデータを書き込む際またはストライプド・ファイルシステムからデータを読み出す際にデータを指定する為のデータファイル名（例えば、ＦＩＬＥＮＡＭＥ１，…）を登録する領域２９１と、そのデータファイル名のデータファイルを複数（ｎ）個の断片に分割した際に各データファイルの断片に割り当てるシーケンス番号（例えば、１，…，ｎ）を登録する領域２９２、及び分割された各データファイルの断片が格納されているファイルシステムのファイルシステム識別子（例えば、ＦＳ－ＩＤ－１１，…，ＦＳ－ＩＤ－１ｎ）を登録する領域２９３とから構成される。

【００１８】図７及び図１０は、ネットワークを構成するホストコンピュータが、広域ファイルシステムを介してデータの入出力を行う際に参照する各種管理テーブルの関係を示した図である。図８は、データの書き込み要求，データの読み出し要求，またはデータの削除要求の時に、ホストコンピュータの間でやり取りされるデータ書き込み要求メッセージ，データ読み出し要求メッセージ，及びデータ削除要求メッセージに含まれるデータの概略を示した図である。データ書き込み要求メッセージ，データ読み出し要求メッセージ，及びデータ削除要求メッセージは、書き込み，読み出し，または削除を行うデータのデータファイル名５２と、データファイルを書き込む，読み出す，または削除すべきデータファイルが格納されているファイルシステム名５３と、データファイルをｎ個に分割した際に各データファイルの断片に割り当てられるシーケンス番号５４とを含んで構成される。図９は、データ書き込み要求メッセージまたはデータ読み出し要求メッセージに応答して、ホストコンピュータの間でやり取りされるデータ書き込み結果通知メッセージ及びデータの読み出し結果通知メッセージに含まれるデータの概略を示した図である。この結果通知メッセージは、書き込みまたは読み出し要求が行われたデータのデータファイル名５２と、データファイルの書き込み要求またはデータファイルの読み出し要求が行われたファイルシステム名５３と、データファイルをｎ個に分割した際に各データファイルの断片に割り当てられたシーケンス番号５４と、データファイルの書き込み要求またはデータファイルの読み出し要求の結果を示す結果フラグ５８と、データ読み出し要求メッセージ受信時に読み出したデータを送り返す為のデータ部５９とを含んで構成される。結果フラグ５８は、成功フラグ（データの書き込みまたはデータの読み出しが成功したことを示すフラグ）及び失敗フラグ（データの書き込みまたはデータの読み出しが失敗したことを示すフラグ）の２つのフラグのうちいずれかを取る。

【００１９】図１１は、図１のシステムの変形例を示す概略システム構成図である。図１１において、コンピュータネットワーク９９で相互に接続された２台のホストコンピュータ９７，９８の各々の２つの入出力制御装置１００，１０１，１０２，１０３に接続された二次記憶装置１０４，１０５，１０６，１０７上に構築されたファイルシステムからストライプド・ファイルシステム１０８が構成されている。ストライプド・ファイルシステム１０８に対するデータ処理は、複数台のホストコンピュータ９７，９８の各々に接続された異る入出力制御装置１００～１０３を介して接続された二次記憶装置１０４～１０７に対し、分散・並列化させて行われる。

【００２０】図１２は、本発明の第２の実施形態例を示す概略システム構成図である。図１３は、図１２のシステムの変形例を示す概略システム構成図である。

【００２１】図１４，１５，１６は本発明における処理の流れを示したフローチャートである。図１４は、ネットワークを構成するホストコンピュータが、広域ファイルシステムを介してデータの書き込みを行う際の処理の流れを示す。図１５は、ネットワークを構成するホストコンピュータ上で動作するプログラムが、遠隔のホストコンピュータからのデータ書き込み要求メッセージ，またはデータ読み出し要求メッセージを受け取った時の処理の流れを示す。図１６はネットワークを構成するホストコンピュータが、広域ファイルシステムを介してデータの読み出しを行う際の処理の流れを示す。

【００２２】次に，上記図面を参照して本発明の動作について詳細に説明する。最初に図１に示す第１の実施形態例について説明するが、その前に本発明の特徴を分かり易くするために、従来から用いられている技術を説明する。

【００２３】従来の広域ファイルシステムにおいては、広域ファイルシステムを管理するための情報としてホスト名，ファイルシステム名，広域ファイルシステム名，及びネットワーク内でホストコンピュータを一意に識別するためのネットワークアドレス情報しか持っていなかった。ストライプド・ファイルシステムを構成している広域ファイルシステムに対する書き込みや読み出し、及びストライプド・ファイルシステムの構成制御は、上記の管理情報をネットワーク内の全てのホストコンピュータ（１１～１５）で共有し、この管理情報を基にして各ホストコンピュータ上のプログラムによってそれぞれ行なわれる。この管理情報の更新は、ホストコンピュータ（１１）においてコマンドによりオペレータから構成変更を指示された場合に行なわれる。一般に、広域ファイルシステムでは、広域ファイルシステム名をネットワークを構成する全てのホストコンピュータで共有して、広域ファイルシステム名からネットワーク内のホストコンピュータの名前とそのホストコンピュータに接続されている二次記憶装置上に構築されているファイルシステム名を参照して、そのホストコンピュータに接続されている二次記憶装置上に構築されているファイルシステムの

ファイルに対して入出力を行う。

【0024】本発明において、広域ファイルシステムを介して異なるホストコンピュータに接続された複数台の二次記憶装置に対して並列にデータの書き込みを行う場合は、図7に示すように、あるホストコンピュータ19からネットワーク10を通して、ホストコンピュータ12～1mのうちのn台のホストコンピュータ1x～1zに接続された二次記憶装置2x～2z上に構築されたファイルシステムから構成された広域ファイルシステム30（例えば、広域ファイルシステム名31＝WA－FS－1n）に対してデータ（データファイル45）の書き込み要求を出す（図14のステップS101）ことによって達成される。

【0025】ホストコンピュータ19上で動作するプログラムは、広域ファイルシステム30に対するデータの書き込み要求を検出すると，その広域ファイルシステム名31（WA－FS－1n）をキーにして、広域ファイルシステム管理用のホストコンピュータ11上に作成されていてネットワーク内の全てのホストコンピュータで共有されているストライプド・ファイルシステム管理テーブル27を検索し、その広域ファイルシステム名31（WA－FS－1n）の広域ファイルシステム30がn個のファイルシステムから構成されていることを知り、書き込もうとしているデータファイル45（例えば、データファイル名46＝FILENAME1）を均等な大きさにn等分して、それぞれ順番に、1からnまでのシーケンス番号47と個別のファイルシステム識別子48（例えば、FS－ID－1，…，FS－ID－n）とを割り当てて、これらデータファイル名46，シーケンス番号47，ファイルシステム識別子48をデータ管理テーブル29に格納する。ホストコンピュータ19上で動作するプログラムは、次に、ストライプド・ファイルシステム管理テーブル27を参照して、指定された広域ファイルシステム30（WA－FS－1n）に割り付けられているn個のファイル識別子32，…，33（例えば、FS－ID－1，…，FS－ID－n）の情報を得る。ファイルシステム識別子32～33の情報を取り出したら、次に広域ファイルシステム管理テーブル26を参照して、ファイルシステム識別子32，…，33に対応するホスト名35，…，36（例えば、Host－1，…，Host－n）及びファイルシステム名37，…，38（例えば、FS－1，…，FS－n）を参照する。ホスト名35，…，36の情報を得たら、アドレス管理テーブル28を参照してn台のホストコンピュータ1x，…，1zのネットワークアドレス42，…，43（例えば、xxx．xxx．xxx．xxx，…，zzz．zzz．zzz．zzz）を得る。このようにして取り出したネットワークアドレス42～43の各ホストコンピュータ1x～1z上で動作しているプログラムの各々に対して，n等分したデータファイル45の各断片のデータ書き込み要求メッセージRw（図8）を送り（図14のステップS101）、広域ファイルシステム管理テーブル26において、データ書き込み要求メッセージRwを送ったホスト名35～36及びファイルシステム名37～38の状態フラグ39～40にBUSYフラグを設定する。

【0026】ホストコンピュータ1x～1z上で動作している各プログラムは、それぞれ自分に接続されている二次記憶装置2x～2z上に構築されているファイルシステム30x～30z（ファイルシステム名37＝（FS－1）～ファイルシステム名38＝（FS－n））へのデータファイル45の断片の書き込み要求メッセージRwを受け取ったら（図15のステップS201）、それぞれ自分に接続されている二次記憶装置2x～2zに対するデータの書き込み処理を行う（図15のステップS202）。このとき、各プログラムは、それぞれホストコンピュータ19から受け取ったデータ書き込み要求メッセージRwのデータファイル名52（FILENAME1）にシーケンス番号54（1～n）を繋げた名前をデータファイル名として選択し、ファイルシステム名53（FS－1～FS－n）で指定されたファイルシステム（30x～30z）にデータの書き込みを行う。ホストコンピュータ1x～1z上で動作している各プログラムは，データ書き込み要求メッセージRwを受け取ってから二次記憶装置2x～2zへのデータの書き込みがあらかじめ設定された一定時間内に完了した場合（図15のステップS203のY側）、データ書き込み結果通知メッセージQw（図9）の結果フラグ58に成功フラグを設定し、さらにデータ書き込み要求メッセージRwに設定されていたデータファイル名52，ファイルシステム名53及びシーケンス番号54を設定して、データの書き込み要求を行ったホストコンピュータ19に対して送る（図15のステップS204）。

【0027】ホストコンピュータ19上で動作しているプログラムは，広域ファイルシステム30に対するデータ書き込み要求を行ってから（図14のステップS101）、一定時間内にデータ書き込み要求を行った全てのホストコンピュータ1x～1z上で動作している各プログラムからデータ書き込み結果通知メッセージQwを受け取り（図14のステップS102、Nの場合）、且つ受け取った全てのデータ書き込み結果通知メッセージQwの結果フラグ58に成功フラグが設定されていた場合（図14のステップS103、成功の場合）に、ホストコンピュータ1x～1zに接続された二次記憶装置2x～2z上に構築されたファイルシステム30x～30zにデータが書き込まれたことを検出し（図14のステップS104）、広域ファイルシステム管理テーブル26においてデータの書き込みが成功したホスト名35～36及びファイルシステム名37～38の状態フラグ39～40をIDLE状態にする。

【0028】ストライプド・ファイルシステムを構成するn台のホストコンピュータ1x～1z上で動作している各プログラムは、ホストコンピュータ19からのデータ書き込み要求メッセージRwを受け取ってから（図15のステップS201）、二次記憶装置2x～2zへのデータの書き込みが一定時間内に完了しなかった場合（図15のステップS203、Nの場合）、データ書き込み結果通知メッセージQwの結果フラグ58に失敗フラグを設定し、更にデータ書き込み要求メッセージRwに設定されていたデータファイル名52，ファイルシステム名53及びシーケンス番号54をそれぞれデータ書き込み結果通知メッセージQwに設定して、データ書き込み要求を行ったホストコンピュータ19に対して送る（図15のステップS205）。

【0029】ホストコンピュータ19上で動作しているプログラムは、広域ファイルシステム30に対するデータの書き込み要求を行ってから、データ書き込み要求メッセージRwを送った全てのホストコンピュータ1x～1z上で動作している各プログラムから一定時間内に返されたデータ書き込み結果通知メッセージQwの結果フラグ58に成功フラグが設定されているものが1つも無かった場合、またはデータ書き込み要求メッセージを送った全てのホストコンピュータ1x～1z上で動作している各プログラムから一定時間内にデータ書き込み結果通知メッセージQwを受け取ることが出来なかった場合は、ホストコンピュータ1x～1zに接続された二次記憶装置2x～2z上に構築されたファイルシステム30x～30zに対するデータの書き込みが失敗したことを検出し（図14のステップS107）、データ管理テーブル29から、書き込みが失敗したデータファイル名46，書き込みが失敗したデータファイル名46に対応するシーケンス番号47及びファイルシステム識別子48のエントリを削除し、広域ファイルシステム管理テーブル26において、データの書き込みに失敗したホスト名35～36及びファイルシステム名37～38の状態フラグをIDLE状態にし、データ管理テーブル29からデータファイル名46のエントリを削除する（図14のステップS108）。

【0030】次に、ホストコンピュータ19が、広域ファイルシステム30に対するデータの書き込み要求を行ってから（図14のステップS101）、一定時間内に各ホストコンピュータ1x～1z上で動作している各プログラムから受け取ったデータ書き込み結果通知メッセージQwの内，一部のデータ書き込み結果通知メッセージQwの結果フラグ58に失敗フラグが設定されていた場合（図14のステップS103、失敗の場合）の動作を図10を共に参照して説明する。例えばi番目のシーケンス番号76に対して設定されたファイルシステム識別子78（例えば、FS－ID－i）に対応するネットワークアドレス68（例えば、sss．sss．ss s．sss）のホスト名61（例えば、Host－i）のホストコンピュータ1sに接続された二次記憶装置2sに構築されたファイルシステム30s（ファイルシステム名63＝（FS－i））に対する、データファイル45のi番目の断片のデータ書き込み結果通知メッセージQwの結果フラグ58に失敗フラグが設定されていた場合、またはホストコンピュータ1x～1z上で動作している各プログラムの中の一部から一定時間以内にデータ書き込み結果通知メッセージQwを受け取ることができなかった場合（図14のステップS102，Yの場合）、例えばi番目のシーケンス番号76に対して設定されたファイルシステム識別子78に対応するネットワークアドレス68のホスト名61のホストコンピュータ1sに接続された二次記憶装置2sに構築されたファイルシステム名63に対するデータファイル45のi番目の断片のデータ書き込み結果通知メッセージQwを一定時間内に受け取ることが出来なかった場合に、データの書き込みが一部分だけ失敗したことを検出し、広域ファイルシステム管理テーブル26において、データの書き込みに失敗したファイルシステム識別子78の状態フラグ65をIDLE状態にする。

【0031】次に、n個に分割されたデータファイル45のi番目のシーケンス番号76のデータ書き込み要求が失敗した場合、ホストコンピュータ19は、図10のデータ管理テーブル29において、（i＋1）番目のシーケンス番号77に割り付けられているファイルシステム識別子79（FS－ID－（i＋1））をキーにして広域ファイルシステム管理テーブル26を検索して、（i＋1）番目のシーケンス番号77に割り付けられているファイルシステム識別子79の状態フラグ66を調べる。（i＋1）番目のシーケンス番号77に割り付けられているファイルシステム識別子79の状態フラグ66がBUSY状態だった場合は、（i＋2）番目のファイルシステム識別子の状態フラグをチェックする。この手順を状態フラグがIDLE状態のファイルシステム識別子が見つかるまで、n番目のファイルシステム識別子まで繰り返し、n番目のファイルシステム識別子までチェックしてもIDLE状態のファイルシステム識別子が見つからなかった場合は、1番目のファイルシステム識別子から（i－1）番目のファイルシステム識別子まで、IDLE状態のファイルシステム識別子が見つかるまで繰り返す。このようにして、広域ファイルシステム管理テーブル26の全てのファイルシステム識別子をチェックしても、IDLE状態のファイルシステム識別子が見つからなかった場合は（図14のステップS105、Yの場合）、データの書き込み要求を行ったホストコンピュータ19はデータファイル45の書き込みが失敗したことを検出し（図14のステップS107）、n等分されたデータファイルの断片を書き込んだ各ファイルシステム識別子に対応するホストコンピュータに対し

て、それぞれ書き込んだn等分されたデータファイルの断片の削除要求を出して、データ管理テーブル２９からデータファイル４５（データファイル名４６＝ＦＩＬＥＮＡＭＥ１）のエントリを削除する（図１４のステップＳ１０８）。

【００３２】（ｉ＋１）番目のシーケンス番号７７に割り付けられているファイルシステム識別子７９の状態フラグ６６がＩＤＬＥ状態の場合は、データ管理テーブル２９において、ｉ番目のシーケンス番号７６に対するファイルシステム識別子７８に、広域ファイルシステム管理テーブル２６においてＩＤＬＥ状態だった（ｉ＋１番目のシーケンス番号７７のデータを書き込んだファイルシステム識別子７９と同じ値（ＦＳ－ＩＤ－（ｉ＋１））をｉ番目のシーケンス番号７６に対するファイルシステム識別子７８に設定して、ｎ個に分割されたデータファイル４５のｉ番目のシーケンス番号７６のデータの断片を、ファイルシステム識別子７８に対応するネットワークアドレス６９（例えば、ｔｔｔ．ｔｔｔ．ｔｔｔ．ｔｔｔ）のホスト名６２（Ｈｏｓｔ－（ｉ＋１））のホストコンピュータ１ｔに接続された二次記憶装置２ｔに構築されたファイルシステム３０ｔ（ファイルシステム名６４＝（ＦＳ＋（ｉ＋１））に書き込む為のデータ書き込み要求メッセージＲｗを出して、広域ファイルシステム管理テーブル２６の（ｉ＋１）番目のファイルシステム識別子の状態フラグ６６をＢＵＳＹ状態にする。

【００３３】ｎ個に分割されたデータファイル４５のｉ番目のシーケンス番号７６のデータの断片を（ｉ＋１）番目のファイルシステム識別子に対応するホストコンピュータ１ｔに送ったデータ書き込み要求メッセージＲｗに対する、データ書き込み結果通知メッセージＱｗが、一定時間内にホストコンピュータ１ｔからデータ書き込み要求を行ったホストコンピュータ１９に返されて、且つそのデータ書き込み結果通知メッセージＱｗの結果フラグ５８に成功フラグがセットされていた場合は、データ書き込み要求を行ったホストコンピュータ１９はデータの書き込みが成功したことを検出し、（ｉ＋１）番目のファイルシステム識別子の状態フラグ６６をＩＤＬＥ状態にする。この場合、ｎ等分されたデータファイル４５のｉ番目のシーケンス番号７６のデータの断片と、（ｉ＋１）番目のシーケンス番号７７のデータの断片とは、（ｉ＋１）番目のファイルシステム識別子７９に対応する、ネットワークアドレス６９のホスト名６２のホストコンピュータ１ｔに接続された二次記憶装置２ｔに構築されたファイルシステム３０ｔに、各々データファイル名４６（ＦＩＬＥＮＡＭＥ１）にシーケンス番号ｉ及び（ｉ＋１）が付与されたデータファイル名で書き込まれる。

【００３４】ｎ個に分割されたデータファイル４５のｉ番目のシーケンス番号７６のデータの断片を（ｉ＋１）番目のファイルシステム識別子に対応するホストコンピュータ１ｔに送ったデータ書き込み要求メッセージＲｗに対するデータ書き込み結果通知メッセージＱｗが、一定時間内にホストコンピュータ１ｔからデータ書き込み要求を行ったホストコンピュータ１９に返されて、且つそのデータ書き込み結果通知メッセージＱｗの結果フラグ５８に失敗フラグがセットされていた場合、または、（ｉ＋１）番目のファイルシステム識別子に対応するホストコンピュータ１ｔに送ったデータ書き込み要求メッセージＲｗに対するデータ書き込み結果通知メッセージＱｗが、一定時間内にホストコンピュータ１ｔからデータ書き込み要求を行ったホストコンピュータ１９に返されなかった場合は、ホストコンピュータ１９は、前述のシーケンスで（ｉ＋２）番目のファイルシステム識別子に対応するホストコンピュータに接続された二次記憶装置に構築されたファイルシステムに対して、ｎ等分されたデータファイル４５のｉ番目のシーケンス番号７６のデータの断片の書き込みを試みる。この手順を，データ書き込みが成功するまで、ｎ番目のファイルシステム識別子まで繰り返し、ｎ番目のファイルシステム識別子に対する書き込みも失敗した場合は、１番目のファイルシステム識別子から（ｉ－１）番目のファイルシステム識別子まで，データの書き込みが成功するまで繰り返す（図１４のステップＳ１０６）。このようにして，広域ファイルシステム管理テーブル２６の全てのファイルシステム識別子に対応するホストコンピュータに接続された二次記憶装置に構築されたファイルシステムに対してｎ等分されたデータファイル４５の断片の書き込みを試みても成功しなかった場合は（図１４のステップＳ１０５、Ｙの場合）、データの書き込み要求を行ったホストコンピュータ１９はデータファイル４５の書き込みが失敗したことを検出し（図１４のステップＳ１０７）、ｎ等分されたデータファイルの断片を書き込んだ各ファイルシステム識別子に対応するホストコンピュータに対して、書き込んだｎ等分されたデータファイルの断片の削除要求を出して、データ管理テーブルからデータファイル４５のエントリを削除する（図１４のステップＳ１０８）。

【００３５】ホストコンピュータ１９が広域ファイルシステム３０を介してデータファイル名４６のデータファイルに格納されているデータの読み出しを行う場合、ホストコンピュータ１９上で動作するプログラムは、データファイル名４６（ＦＩＬＥＮＡＭＥ１）のデータファイル４５に格納されているデータの読み出し要求を検出すると、データファイル名４６をキーにして、広域ファイルシステム管理用のホストコンピュータ１１上に作成されていてネットワーク内の全てのホストコンピュータで共有されているデータ管理テーブル２９を参照して、データファイル名４６のデータファイルが断片化されている数をシーケンス番号４７から知り、そして各断片化

されたデータファイルが格納されているファイルシステムのファイルシステム識別子48（32～33）を取り出す。次に，各断片化されたデータファイルが格納されているファイルシステムのファイルシステム識別子48（32～33）をキーにして、広域ファイルシステム管理テーブル26を検索し、各ファイルシステム識別子32～33に対応するホスト名35～36、及びファイルシステム名37～38を取り出す。次に、ホスト名35～36をキーにしてアドレス管理テーブル28を参照して、データファイル名46の断片化されたデータファイルが保存されている二次記憶装置2x～2zが接続されているホスト名35～36（ホストコンピュータ1x～1z）のネットワークアドレス42～43を取り出す。このようにして取り出したネットワークアドレス42～43の各ホストコンピュータ1x～1z上で動作しているプログラムの各々に対して、データ読み出し要求メッセージRrを送り（図16のステップS301）、広域ファイルシステム管理テーブル26において、データ読み出し要求メッセージRrを送ったホスト名35～36及びファイルシステム名37～38の状態フラグ39～40にBUSYフラグを設定する。

【0036】ホストコンピュータ1x～1z上で動作している各プログラムは、データ読み出し要求メッセージを受け取ったら（図15のステップS201）、自分に接続されている二次記憶装置2x～2zからのデータの読み出し処理を行う（図15のステップS202））。このとき、ホストコンピュータ1x～1z上で動作している各プログラムは、自分に接続されている二次記憶装置2x～2zからデータを読み出す際に、データファイル名として、ホストコンピュータ19から受け取ったデータ読み出し要求メッセージRrのデータファイル名52にシーケンス番号54を繋げた名前をデータファイル名として選択し、ファイルシステム名53からデータの読み出しを行う。ホストコンピュータ1x～1z上で動作している各プログラムは、データ読み出し要求メッセージRrを受け取ってから、二次記憶装置2x～2zからのデータの読み出しが一定時間内に完了した場合（図15のステップS203、Yの場合）、データ読み出し結果通知メッセージQrの結果フラグ58に成功フラグを設定し、更にデータ読み出し要求メッセージRrに設定されていたデータファイル名52，ファイルシステム名53及びシーケンス番号54を、各々データ読み出し結果通知メッセージQrに設定して、二次記憶装置から読み出したデータファイルの内容をデータ部59に書き込んで、データの読み出し要求を行ったホストコンピュータ19に対して送る（図15のステップS204）。コンピュータ19上で動作しているプログラムは、広域ファイルシステム30に対するデータ読み出し要求を行ってから（図16のステップS301）、一定時間内にデータ読み出し要求メッセージRrを送った全てのホストコンピュータ1x～1z上で動作している各プログラムからデータ読み出し結果通知メッセージQrを受け取り（図16のステップS302、Nの場合）、且つ受け取った全てのデータ読み出し結果通知メッセージQrの結果フラグ58に成功フラグが設定されていた場合（図16のステップS303、成功の場合）に、ホストコンピュータ1x～1zに接続された二次記憶装置2x～2z上に構築されたファイルシステム30x～30z（ファイルシステム名37～38）からのデータの読み出しが成功したことを検出し（図16のステップS304）、広域ファイルシステム管理テーブル26において、データの読み出しが成功したホスト名35～36及びファイルシステム名37～38の状態フラグ39～40をIDLE状態にする。コンピュータ19上で動作しているプログラムは、データ読み出し要求メッセージRrで指定したデータファイル45の読み出しが成功したことを検出すると、各ホストコンピュータから送られて来たデータ読み出し結果通知メッセージQrからデータ部59を取り出し、シーケンス番号57の順番に並べ換えた上で、元の1つのファイルに組み立てる（図16のステップS305）。このようにして，ネットワーク内の複数のホストコンピュータに接続された二次記憶装置に分割して保存されていたデータファイル45にアクセスできるようになる。

【0037】ストライプド・ファイルシステムを構成するn台のホストコンピュータ1x～1z上で動作している各プログラムは、ホストコンピュータ19からのデータ読み出し要求メッセージRrを受け取ってから（図15のステップS201）、二次記憶装置2x～2zからのデータの読み出しが一定時間内に完了しなかった場合（図15のステップS203、Nの場合）、データ読み出し結果通知メッセージQrの結果フラグ58に失敗フラグを設定し、更にデータ読み出し要求メッセージRrに設定されていたデータファイル名52，ファイルシステム名53及びシーケンス番号54を、各々データ読み出し結果通知メッセージQrに設定して、データ読み出し要求を行ったホストコンピュータ19に対して送る（図15のステップS205）。

【0038】ホストコンピュータ19上で動作しているプログラムは、広域ファイルシステム30からのデータの読み出し要求を行ってから、データ読み出し要求メッセージRrを送った全てのホストコンピュータ1x～1z上で動作している各プログラムから一定時間内に返されたデータ読み出し結果通知メッセージQrの結果フラグ58に失敗フラグが設定されているものが1つでもあった場合（図16のステップS303、失敗の場合）、またはデータ読み出し要求メッセージRrを送ったホストコンピュータ1x～1z上で動作している各プログラムのうち、一定時間内にデータ読み出し結果通知メッセージQrを返さないものが1つでもあった場合（図16

のステップS302、Yの場合）は、ホストコンピュータ1x～1zに接続された二次記憶装置2x～2z上に構築されたファイルシステム30x～30zからのデータの読み出しが失敗したことを検出し（図16のステップS306）、広域ファイルシステム管理テーブル26において、データの読み出しに失敗したホスト名35～36及びファイルシステム名37～38の状態フラグ39～40をIDLE状態にして、ホストコンピュータ19における広域ファイルシステム30からのデータファイル45の読み出し要求を失敗とする。

【0039】次に、本発明の第2の実施の形態例について図12を参照して説明する。図12において、2つのホストコンピュータ110，111は、コンピュータネットワーク109によって相互に接続されていて、各ホストコンピュータ110，111には、それぞれ2つの入出力制御装置112～113，114～115が搭載されている。各入出力制御装置112，113，114，115には、それぞれ二次記憶装置116，117，118，119が接続されている。第1のストライプド・ファイルシステム120は、ホストコンピュータ110の入出力制御装置112に接続された二次記憶装置116，及びホストコンピュータ111の入出力制御装置114に接続された二次記憶装置118から構成され、第2のストライプド・ファイルシステム121は、ホストコンピュータ110の入出力制御装置113に接続された二次記憶装置117，及びホストコンピュータ111の入出力制御装置115に接続された二次記憶装置119から構成されている。広域ファイルシステムを介したストライプド・ファイルシステム120，121へのデータの入出力を行った場合の二次記憶装置116～119に対するデータの入出力処理は、前述の第1の実施の形態例と同じ手順で行われる。第2の実施形態例において、ホストコンピュータ110が障害により停止した場合、ストライプド・ファイルシステム120，121に対して書き込んだデータは、自動的に、停止していないホストコンピュータ111に搭載されている入出力制御装置114，115に接続された二次記憶装置118，119に保存される。また、第2の実施形態例においてホストコンピュータ110に搭載されている入出力制御装置112が障害により停止した場合、ストライプド・ファイルシステム120に対して書き込んだデータは、自動的に、ホストコンピュータ111に搭載されている停止していない入出力制御装置114に接続された二次記憶装置118に保存される。更に、第2の実施形態例において、ホストコンピュータ110に搭載されている入出力制御装置112に接続されている二次記憶装置116が障害により停止した場合も、ストライプド・ファイルシステム120に対して書き込んだデータは、自動的に、ホストコンピュータ111に搭載されている入出力制御装置114に接続された停止していない二次記憶装置118に保存される。

【0040】更に、第2の実施形態例の変形例を図13に示す。図13において、3つのホストコンピュータ123，124，125はコンピュータネットワーク122によって相互に接続されていて、各ホストコンピュータ123，124，125には、それぞれ2つの入出力制御装置126～127，128～129，130～131が搭載されている。ホストコンピュータ123～125の各々の入出力制御装置126，127，128，129，130，131には、それぞれ二次記憶装置132，133，134，135，136，137が接続されている。第1のストライプド・ファイルシステム138は、ホストコンピュータ123の入出力制御装置126に接続された二次記憶装置132と、ホストコンピュータ124の入出力制御装置128に接続された二次記憶装置134と、ホストコンピュータ125の入出力制御装置130に接続された二次記憶装置136とから構成される。第2のストライプド・ファイルシステム139は、ホストコンピュータ123の入出力制御装置127に接続された二次記憶装置133と、ホストコンピュータ124の入出力制御装置129に接続された二次記憶装置135と、ホストコンピュータ125の入出力制御装置131に接続された二次記憶装置137とから構成される。広域ファイルシステムを介したストライプド・ファイルシステム138，139へのデータの入出力を行った場合の二次記憶装置132～137に対するデータの入出力処理は、前述の第1の実施形態例と同じ手順で行われる。この第2の実施形態例の変形例において、ホストコンピュータ123が障害により停止した場合、ストライプド・ファイルシステム138，139に対してデータの書き込みを行ったプロセスの処理は、自動的に、停止していないホストコンピュータ124，125に搭載されている入出力制御装置128～131により分散・並列処理され、データはホストコンピュータ124，125に搭載されている入出力制御装置128，130に接続された二次記憶装置134，136に保存される。また、ホストコンピュータ123に搭載されている入出力制御装置126が障害により停止した場合、ストライプド・ファイルシステム138に対してデータの書き込みを行ったプロセスの処理は、ホストコンピュータ124，125に搭載されている入出力制御装置128，130により分散・並列処理され、データはホストコンピュータ124，125に搭載されている入出力制御装置128，130に接続された二次記憶装置134，136に保存される。さらにまた、ホストコンピュータ123に搭載されている入出力制御装置126に接続されている二次記憶装置132が障害により停止した場合も、ストライプド・ファイルシステム138に対してデータの書き込みを行ったプロセスの処理は、ホストコンピュータ124，125に搭載されている入出

力制御装置１２８，１３０により分散・並列処理され、データはホストコンピュータ１２４，１２５に搭載されている入出力制御装置１２８，１３０に接続された二次記憶装置１３４，１３６に保存される。

【００４１】以上説明したように、本発明によれば、ストライプド・ファイルシステムに対する入出力処理を，コンピュータネットワーク内の複数のホストコンピュータに分散・並列化させることができるようになるので、従来の１台のホストコンピュータに接続された複数台の二次記憶装置上に構築されたファイルシステムからストライプド・ファイルシステムを構築した場合（図１７参照）に１台のホストコンピュータ８０で全ての二次記憶装置８１～８４に対するデータの入出力を処理しなければならなかったのに比べて、同じ量のデータの入出力を複数台のホストコンピュータ（図1の１２～１ｍ）で分割して並行に処理することができるようになるため、各ホストコンピュータが処理しなければならない入出力データの量が減少する。例えば、従来のストライプド・ファイルシステムの手法では、１台のホストコンピュータが２つの入出力制御装置を搭載し、その各々の入出力制御装置に二次記憶装置が接続されていた場合（図１８参照）、プロセスが二次記憶装置８９，９０に対するデータの入出力処理は２つの入出力制御装置８７，８８によって分散・並列処理されるに過ぎないが、本発明によれば、コンピュータネットワークに接続され２つの入出力制御装置を搭載し、その各々の入出力制御装置に二次記憶装置が接続されているもう１台のホストコンピュータとの間でストライプド・ファイルシステムを構築することにより（図１１参照）、二次記憶装置に対するデータの入出力処理を２台のホストコンピュータに接続された４つの入出力制御装置によって分散・並列化して処理することが可能となり、二次記憶装置に対するデータの入出力時に、各々のホストコンピュータが処理しなければならない入出力データの量を減少させることができる。このように、二次記憶装置に対する入出力処理そのものをコンピュータネットワーク内の複数のホストコンピュータ上で動作する複数のプロセスに分散させ、各ホストコンピュータにより複数の入出力制御装置に分散・並列化させることにより、二次記憶装置に対するデータの入出力時に各ホストコンピュータにかかる負荷を軽減することができるようになる。

【００４２】また、従来のストライプド・ファイルシステムの手法では、複数台の二次記憶装置に対するデータの入出力を１台の入出力制御装置で行った場合は（図１９参照）、ストライプド・ファイルシステム９６に対する入出力処理を並列化しても、実際の二次記憶装置９４，９５へのデータの入出力処理は１台の入出力制御装置９３によってシーケンシャルに行われる為、単一プロセス当たりの二次記憶装置に対する入出力時の性能向上を図ることができず、二次記憶装置に対するデータの入出力時のプロセス当たりの性能向上を図るには（図１８参照）、異なる入出力制御装置８７，８８に接続された二次記憶装置８９，９０上に構築されたファイルシステムを基本単位としてストライプド・ファイルシステム９１を構築しなければならないというハードウェア上の制約があった。この為、ストライプド・ファイルシステムを構築し二次記憶装置に対するデータの入出力処理を複数の入出力制御装置に分散・並列化させることによりプロセス当たりの性能向上を図った場合、性能向上を実現することができるストライプド・ファイルシステムの大きさは１台のホストコンピュータに搭載されている入出力制御装置の数によって制限されていたが、本発明によれば、１台のホストコンピュータの複数の異なる入出力制御装置に接続された二次記憶装置だけでなく、コンピュータネットワークを構成する複数台のホストコンピュータに搭載された複数の異なる入出力制御装置に接続された二次記憶装置からストライプド・ファイルシステムを構築できるようになる。例えば、従来のストライプド・ファイルシステムの手法では、１台のホストコンピュータに２つの入出力制御装置が搭載され、その各々の入出力制御装置に二次記憶装置が接続されていた場合（図１８）、二次記憶装置に対するデータの入出力処理を分散・並列化することにより，二次記憶装置８９，９０に対するデータの入出力時の性能向上を図ることができるストライプド・ファイルシステムの容量は、２つの入出力制御装置８７，８８の各々に接続された二次記憶装置８９，９０の容量の合計にしかならない。本発明によれば、図１１に示すようにコンピュータネットワーク９９により接続され、２つの入出力制御装置を搭載し、その各々の入出力制御装置に二次記憶装置が接続されているもう１台のホストコンピュータとの間でストライプド・ファイルシステム１０８を構築することにより、２台のホストコンピュータ９７，９８の４つの入出力制御装置１００～１０３の各々に接続された二次記憶装置１０４～１０７の容量の合計になり、二次記憶装置に対するデータの入出力処理を分散・並列化することによる性能向上を犠牲にすることなく、１台のホストコンピュータだけでストライプド・ファイルシステムを構築して二次記憶装置に対するデータの入出力処理を分散・並列化した場合（図１８）に比べて２倍の容量のストライプド・ファイルシステムを構築できるようになる。このように、コンピュータネットワークを経由して複数台のホストコンピュータに接続された二次記憶装置上に構築された広域ファイルシステムを構成単位としたストライプド・ファイルシステムを構築することにより、二次記憶装置に対するデータの入出力時のプロセス当たりの性能向上を図りつつ、大容量のストライプド・ファイルシステムを構築できるようになる。

【００４３】さらに本発明によれば、ストライプド・ファイルシステムを介してプロセスが二次記憶装置にデー

タを書き込んだ場合、ストライプド・ファイルシステムを構成しているファイルシステムが構築されている二次記憶装置が接続されているホストコンピュータのうちの1台が障害により停止しても、同じストライプド・ファイルシステムを構成しているファイルシステムが構築されている二次記憶装置が接続されている別のホストコンピュータが動作していて、その二次記憶装置にデータを保存するだけの容量がある限り、プロセスがストライプド・ファイルシステムに書き込んだデータは自動的に正常に動作しているホストコンピュータに接続されている二次記憶装置に保存される。このように、ストライプド・ファイルシステムに対してデータの入出力を行う際、コンピュータネットワークを介して異なるホストコンピュータに接続された二次記憶装置上に構築されたファイルシステムに同時にアクセス可能とすることにより、コンピュータネットワーク全体で考えた場合のシステムの耐故障性を向上させることができるようになる。

【0044】

【発明の効果】以上説明したように、本発明によれば、二次記憶装置に対する入出力処理そのものをコンピュータネットワーク内の複数のホストコンピュータ上で動作する複数のプロセスに分散させ、各ホストコンピュータにより複数の入出力制御装置に分散・並列化させることにより、二次記憶装置に対するデータの入出力時に各ホストコンピュータにかかる負荷を軽減することができる。

【0045】また、コンピュータネットワークを経由して複数台のホストコンピュータに接続された二次記憶装置上に構築された広域ファイルシステムを構成単位としたストライプド・ファイルシステムを構築することにより、二次記憶装置に対するデータの入出力時のプロセス当たりの性能向上を図りつつ、大容量のストライプド・ファイルシステムを構築できる。

【0046】さらに、ストライプド・ファイルシステムに対してデータの入出力を行う際、コンピュータネットワークを介して異なるホストコンピュータに接続された二次記憶装置上に構築されたファイルシステムに同時にアクセスすることにより、コンピュータネットワーク全体で考えた場合のシステムの耐故障性を向上させることができる。

【図面の簡単な説明】

【図1】本発明の第1の実施の形態例を示すシステムの構成図である。

【図2】広域ファイルシステムを管理する為のホストコンピュータに接続されている二次記憶装置上の各種管理テーブルを示す図である。

【図3】広域ファイルシステム管理テーブルの構成例を示す図である。

【図4】ストライプド・ファイルシステム管理テーブルの構成例を示す図である。

【図5】アドレス管理テーブルの構成例を示す図である。

【図6】データ管理テーブルの構成例を示す図である。

【図7】本発明においてデータの書き込み及び読み出しを行う際に、参照する各種管理テーブルの関係を示した図である。

【図8】データ書き込み要求メッセージ及びデータ読み出し要求メッセージの構成例を示す図である。

【図9】データ書き込み結果通知メッセージ及びデータ読み出し結果通知メッセージの構成例を示す図である。

【図10】データの書き込み要求が失敗した場合のリトライ処理の際に，参照する各種管理テーブルの関係を示した図で、図7に対する補足図である。

【図11】図1に示す実施の形態の変形例を示す図である。

【図12】本発明の第2の実施の形態例を示すシステムの構成図である。

【図13】図12に示す実施の形態の変形例を示す図である。

【図14】本発明において、ホストコンピュータが広域ファイルシステムを介してデータの書き込みを行う際の処理の流れを示したフローチャートである。

【図15】本発明において、ホストコンピュータ上で動作するプログラムが遠隔のホストコンピュータからのデータの書き込み要求，またはデータの読み出し要求を受け取った時の処理の流れを示したフローチャートである。

【図16】本発明において、ホストコンピュータが広域ファイルシステムを介してデータの読み出しを行う際の処理の流れを示したフローチャートである。

【図17】従来のストライプド・ファイルシステムを示す構成図である。

【図18】従来のストライプド・ファイルシステムにおいて、1台のホストコンピュータの異なる入出力制御装置に接続された二次記憶装置に対するデータの入出力処理を、分散・並列化させて行う場合のハードウェア構成を示す図である。

【図19】従来のストライプド・ファイルシステムにおいて、1台のホストコンピュータに接続された複数の二次記憶装置に対するデータの入出力処理を1つの入出力制御装置により行う場合のハードウェア構成を示す図で、図18に対する補足図である。

【符号の説明】

10，99，109，122　ホストコンピュータを相互に接続するネットワーク

11　広域ファイルシステム管理用のホストコンピュータ

12～1m，97，98，110，111，123，124，125　ネットワークを構成するホストコンピュータ

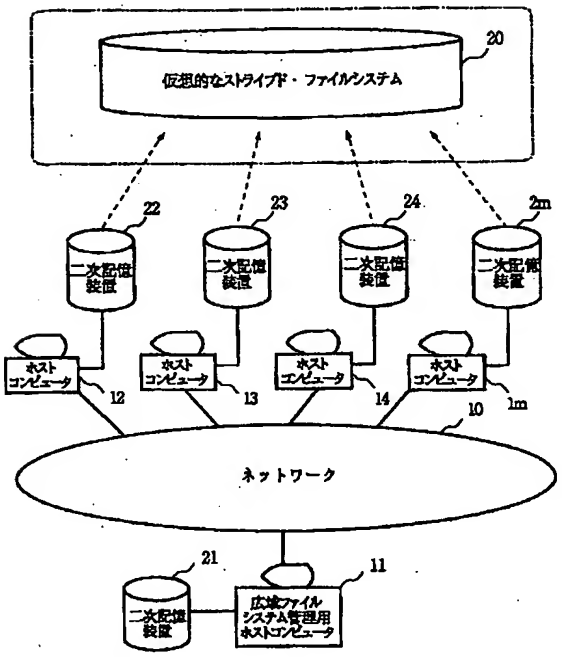
23

２１～２ｍ，１０４～１０７，１１６～１１９，１３２～１３７　二次記憶装置
２０，１０８，１２０，１２１，１３８，１３９　仮想的なストライプド・ファイルシステム
２６　広域ファイルシステム管理テーブル
２７　ストライプド・ファイルシステム管理テーブル
２８　アドレス管理テーブル

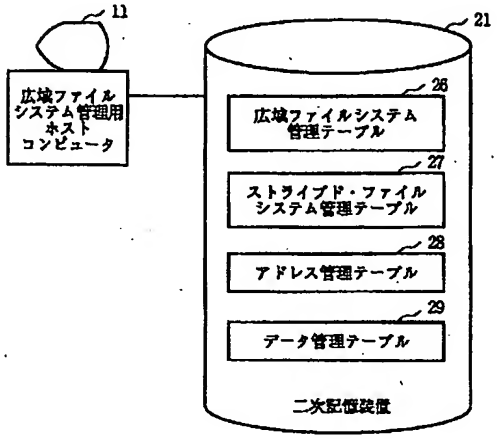
24

＊２９　データ管理テーブル
３０　広域ファイルシステム
３０ｓ～３０ｚ　ファイルシステム
４５　データファイル
１００～１０３，１１２～１１５，１２６～１３１　入出力制御装置

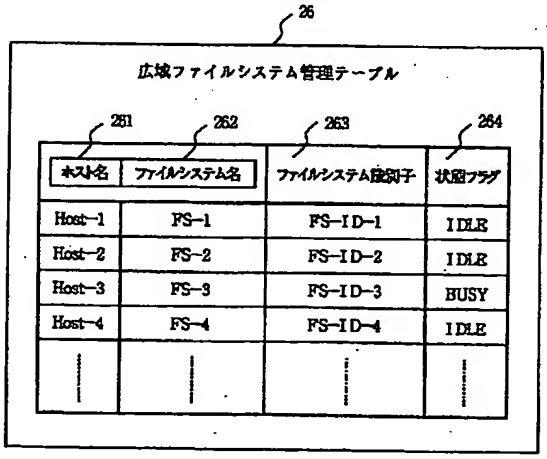
【図１】



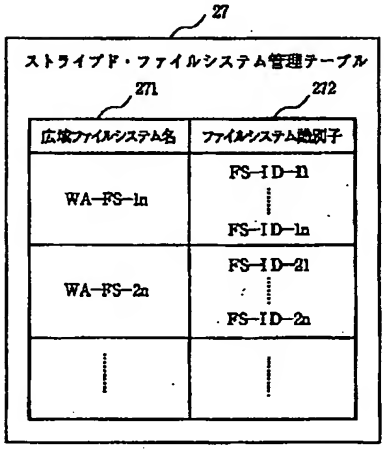
【図２】



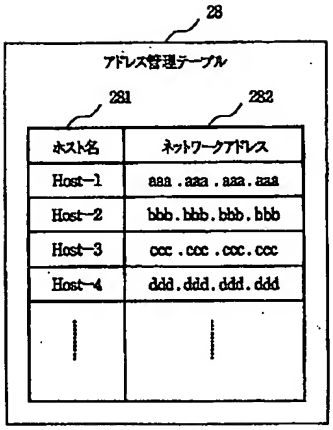
【図３】



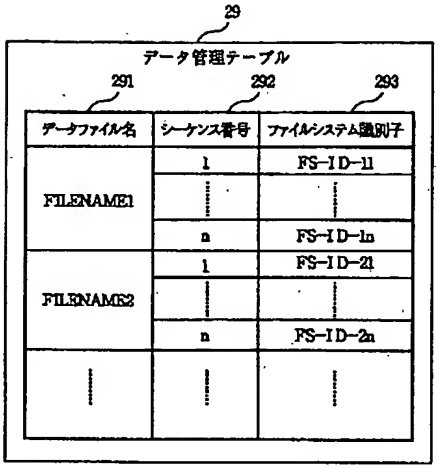
【図４】



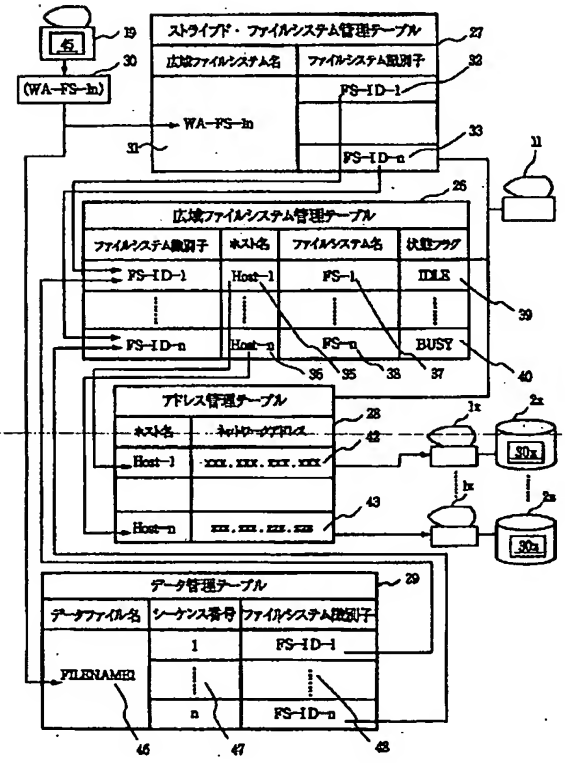
【図5】



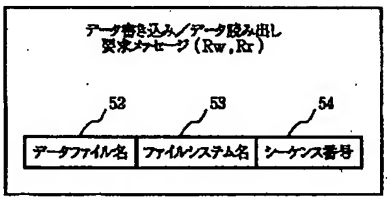
【図6】



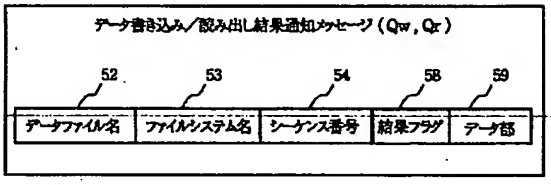
【図7】



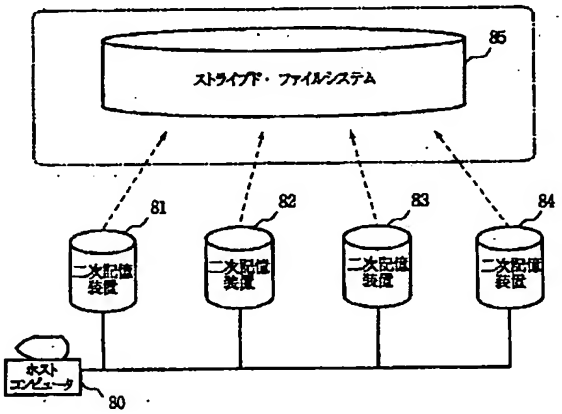
【図8】



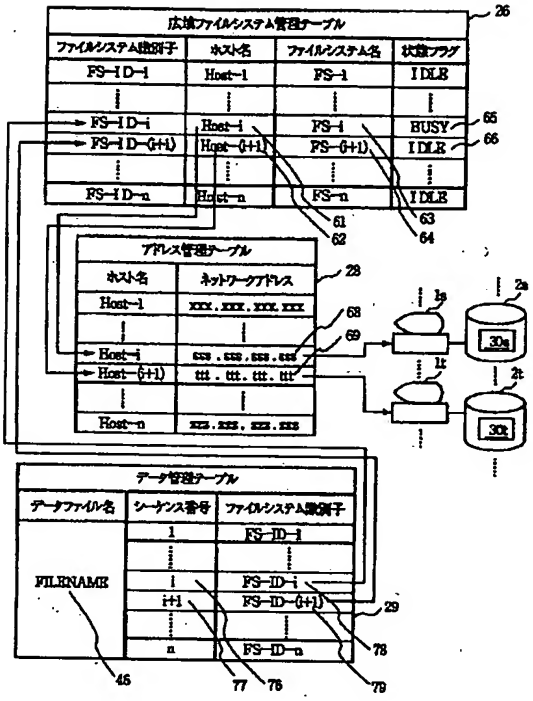
【図9】



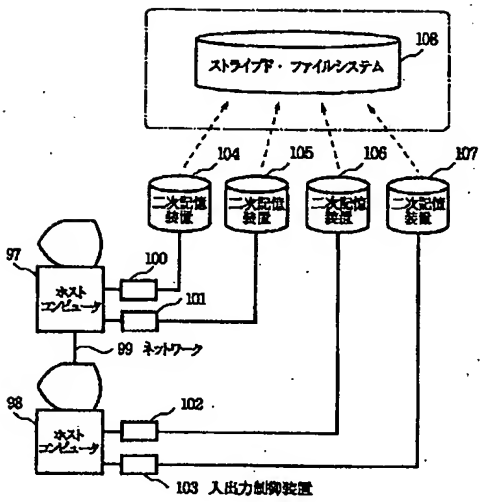
【図17】



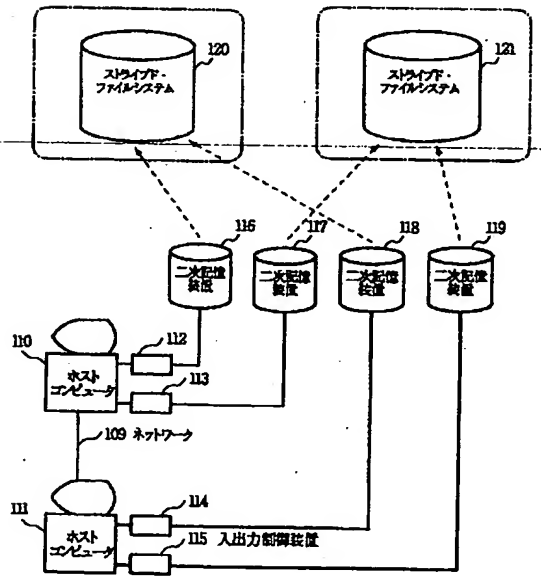
【図１０】



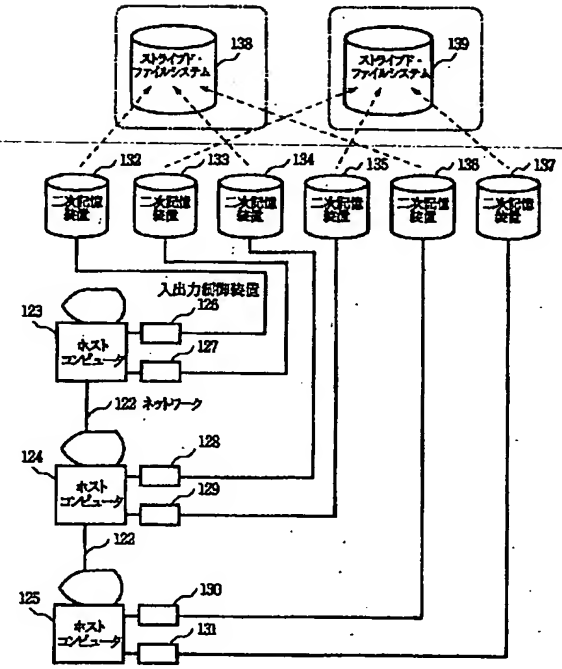
【図１１】



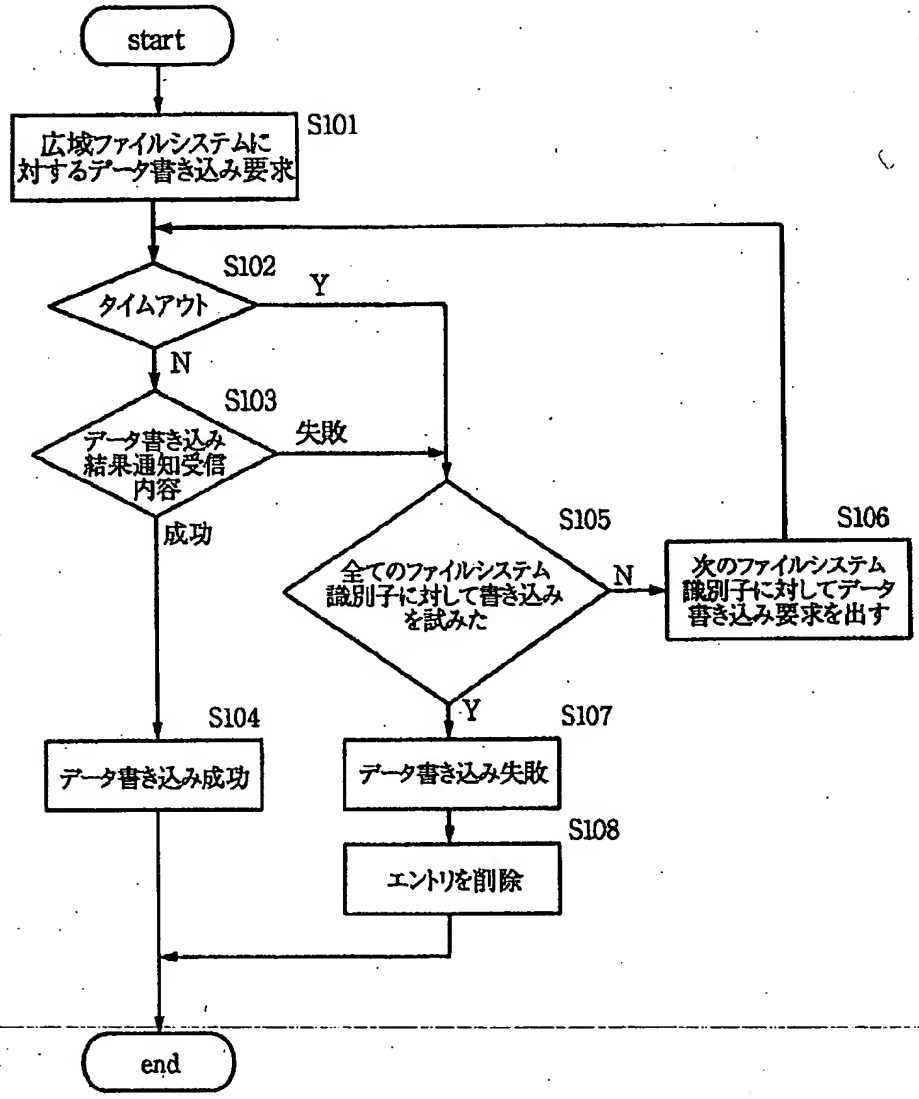
【図１２】



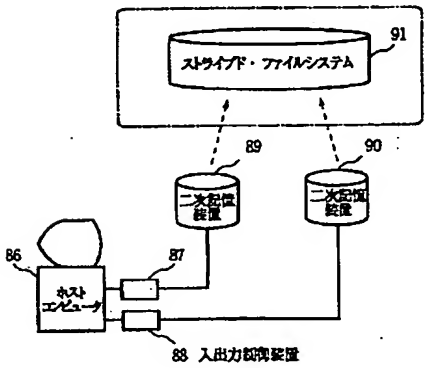
【図１３】



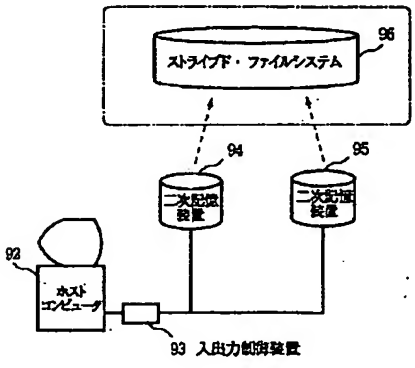
【図１４】


start
広域ファイルシステムに対するデータ書き込み要求
S101
S102
タイムアウト
Y
N
S103
データ書き込み結果通知受信内容
失敗
成功
S105
全てのファイルシステム識別子に対して書き込みを試みた
N
S106
次のファイルシステム識別子に対してデータ書き込み要求を出す
Y
S107
S104
データ書き込み成功
データ書き込み失敗
S108
エントリを削除
end

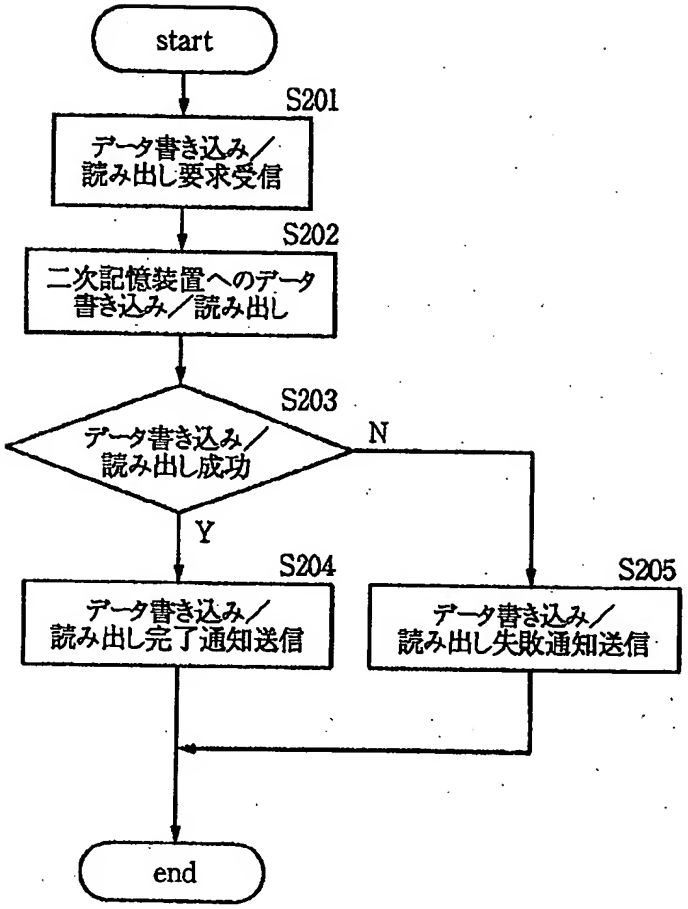
【図１８】


91
ストライプド・ファイルシステム
89
90
二次記憶装置
二次記憶装置
86
ホストコンピュータ
87
88 入出力制御装置

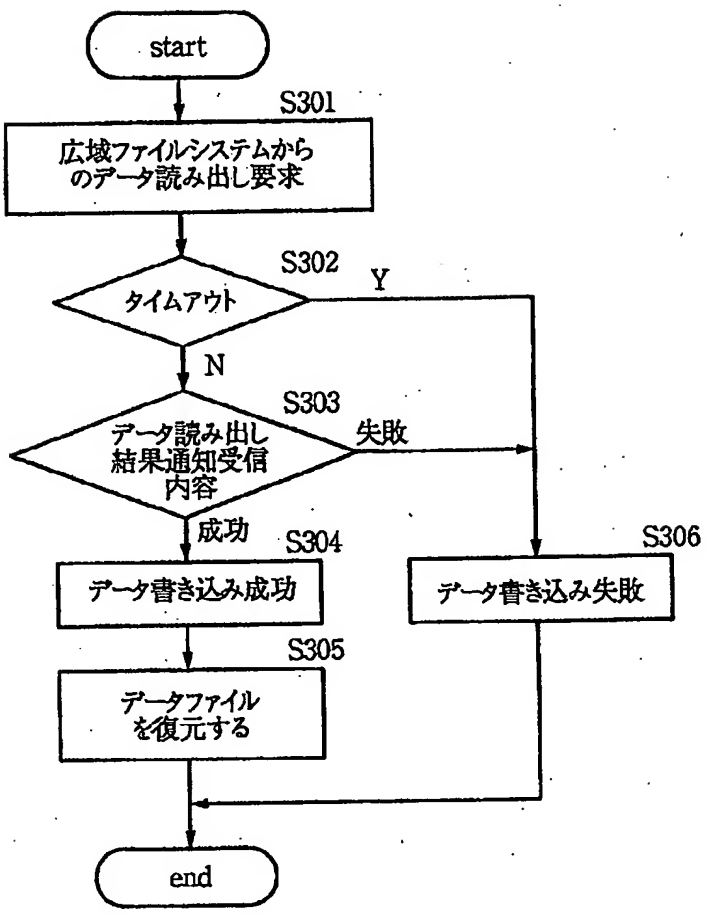
【図１９】


96
ストライプド・ファイルシステム
94
95
二次記憶装置
二次記憶装置
92
ホストコンピュータ
93 入出力制御装置

【図15】


start
S201
データ書き込み／
読み出し要求受信
S202
二次記憶装置へのデータ
書き込み／読み出し
S203
データ書き込み／
読み出し成功
N
Y
S204
データ書き込み／
読み出し完了通知送信
S205
データ書き込み／
読み出し失敗通知送信
end

【図16】


start
S301
広域ファイルシステムから
のデータ読み出し要求
S302
タイムアウト
Y
N
S303
データ読み出し
結果通知受信
内容
失敗
成功
S304
データ書き込み成功
S305
データファイル
を復元する
S306
データ書き込み失敗
end